

Archivage numérique de collections photographiques

—

Rapport de principe



Date : avril 2002

Etabli pour Office fédéral de la protection civile
Section Protection des biens culturels
Monsieur R. Büchel
Monbijoustrasse 91
3003 Berne

De : Rudolf Gschwind
Lukas Rosenthaler
Abt. Bild- und Medientechnologie
Universität de Bâle
Bernoullistrasse 32
4056 Bâle

Franziska Frey
College of Imaging Arts and Sciences
Rochester Institute of Technology
Rochester, NY

1. Introduction.....	3
2. Directives pour l'archivage numérique à long terme	4
2.1. Introduction.....	4
2.2. Critères de procédure.....	6
2.3. Critères techniques.....	9
3. Métadonnées.....	18
3.1. Différentes formes de métadonnées.....	18
3.2. Principes pour l'archivage de métadonnées.....	18
3.3. Différents standards.....	18
3.4. Métadonnées descriptives.....	19
3.4.1. Données de classement/métadonnées administratives:.....	20
3.4.2. Métadonnées techniques.....	20
3.5. Références.....	26
4. Directives pour la numérisation de photographies.....	27
4.1. Introduction.....	27
4.2. Numérisation.....	28
4.2.1. Choix du matériel.....	28
4.2.2. Définition spatiale.....	29
4.2.3. Valeur tonale.....	29
4.2.4. Calibrage et reproduction des couleurs.....	31
4.2.5. Courant d'obscurité, bruit (bruit) et sensibilité	33
4.2.6. Distorsions géométriques.....	34
4.3. Contrôle qualité.....	34
5. Annexe.....	37
5.1. Exemple d'archivage à long terme	37
5.1.1. Numérisation et contrôle qualité.....	37
5.1.2. Migration	38
5.2. Schéma type pour le processus de numérisation de matériels photographiques.....	39

1. Introduction

Le présent rapport comporte des directives pour l'archivage à long terme en toute sécurité de photographies sous la forme d'enregistrements numériques. Il s'articule en trois parties:

- **Directives pour l'archivage numérique à long terme**
Cette partie décrit les conditions à remplir par une archive de données pour que les images numériques restent accessibles aux générations futures.
- **Métadonnées**
Les métadonnées qui comportent les informations sur les données graphiques doivent faire partie intégrante d'une archive de photographies numérique. Cette partie décrit les formes les plus courantes ainsi que les standards actuels en matière de métadonnées.
- **Directives pour la numérisation**
Pour créer une archive d'images numérique, les photos doivent être numérisées dans une qualité adéquate. L'application des directives décrites dans ce document garantit que les copies numériques des photographies originales pourront servir de "substitut numérique" dans la quasi-totalité des applications.

Le rapport se fonde sur le principe selon lequel la copie numérique constitue un substitut à part entière pour toutes les applications où seul le contenu visuel de l'image compte.

Pour l'archivage à long terme, l'archive numérique doit offrir une sécurité au moins équivalente sinon supérieure à la méthode analogique "traditionnelle". En effet, même les archives traditionnelles avec copie de sécurité sur microfilm n'offrent pas de sécurité à 100%.

L'aspect des métadonnées est souvent sous-estimé. De nombreuses collections photographiques sont mal explorées et, si elles existent, les métadonnées ne sont bien souvent ni systématiques, ni organisées (simples notes manuscrites au dos des photos, par exemple). Or l'organisation de ces métadonnées représente un travail considérable, au même titre d'ailleurs que la numérisation de ces données.

La numérisation pour sa part se fonde sur le principe voulant que les copies numériques puissent offrir toutes les applications possibles avec l'image originale, voire davantage (agrandissement de détails, etc.). On pourra parler de "fac-similé numérique" lorsque cette condition sera remplie.

Les directives décrites in extenso dans ce rapport permettent la réalisation avec succès d'une archive numérique à long terme pour des photographies.

2. Directives pour l'archivage numérique à long terme

2.1. Introduction

Deux risques majeurs doivent être réduits au minimum lors de l'archivage durable de données analogiques (notamment de photographies): 1) la destruction totale de l'image par exemple par le feu ou l'eau et 2) la décomposition intrinsèque due au vieillissement naturel du support qui doit être freinée au maximum. Ces deux risques peuvent toutefois être optimisés par un stockage approprié (lieu d'entreposage, protection contre l'incendie, climat contrôlé). Par voie de conséquence, l'accessibilité des données ainsi archivées est fort restreinte et la décomposition intrinsèque peut tout au plus être ralentie mais non stoppée. Et comme chaque opération de copie de données analogiques (photographies, films, vidéos, etc.) entraîne une déperdition de qualité parfois importante, il est indispensable de conserver l'original pour la postérité. Les données analogiques ont souvent la propriété de pouvoir être interprétées sans outils techniques particuliers comme c'est le cas de la photographie, par exemple. Les métadonnées, telles que l'origine de l'image, son auteur, la description de l'objet peuvent souvent être consignées sur le média même (par exemple sous la forme de notes au dos d'une photographie).

Les données numériques ont quant-à elles des propriétés bien différentes des données analogiques pour l'archivage durable:

- Les données numériques peuvent être copiées indéfiniment à condition d'employer le bon procédé. L'original et la copie sont identiques et ne peuvent être distingués, ce qui enlève tout son sens à la notion d'"original" dans le domaine du numérique.
- Le problème de la décomposition continue ne se pose pas avec le numérique. Soit les informations enregistrées sous forme numérique peuvent être lues intégralement (et correctement), soit des erreurs se produisent qui dévalorisent en principe tout l'enregistrement concerné. Des algorithmes spéciaux ("sommets de contrôle") doivent donc être mis en place pour détecter la survenue d'erreurs.¹

Quels sont les mécanismes susceptibles de causer une déperdition d'informations avec les données numériques? Le bref aperçu qui suit portera plus particulièrement sur les problèmes de l'archivage durable dans lequel il faut non seulement préserver l'objet en tant que tel mais aussi les informations sur l'objet ou métadonnées². Une difficulté supplémentaire se pose pour les archives numériques: Les supports de données eux-mêmes ne peuvent être lus et

¹ Avec les supports de données actuels, le calcul et la vérification des "sommets de contrôle" sont généralement intégrés dans le hardware. En cas d'erreurs de lecture, c'est souvent l'enregistrement entier qui est déclaré illisible et sauté ou l'opération de lecture qui est entièrement interrompue. Il n'existe donc que deux possibilités pour les données numériques: l'enregistrement est lisible et donc "correct" ou il est illisible et perdu.

² Les métadonnées sont définies comme des informations sur le contexte d'un objet. Dans le cas des photographies, il peut par exemple s'agir d'un texte descriptif, du nom des personnes représentées, du lieu et de la date/heure, du photographe mais aussi d'informations techniques comme le type de film, le temps d'exposition, etc. Vous trouverez des informations complémentaires au chapitre "Métadonnées".

interprétés qu'avec des outils techniques. Or l'aspect extérieur d'une bande informatique est toujours le même, qu'elle soit vide ou qu'elle contienne des images, du texte ou d'autres données. Une armoire contenant 10 000 cédéroms non marqués n'aura ainsi guère de valeur sans informations complémentaires. Bref, une stratégie inappropriée d'archivage de données numériques peut entraîner une perte d'informations à plusieurs niveaux. Il suffit en effet que des informations se soient perdues à un seul de ces niveaux pour que l'intérêt et la valeur d'une archive numérique soient considérablement altérés, voire totalement détruits.

Il faut distinguer 6 niveaux auxquels des informations peuvent être perdues:

- **Données de classement/métadonnées administratives:** L'information recherchée est introuvable parce qu'il n'y a pas de données de classement. Nous désignons par données de classement toutes les informations servant à retrouver et à lire les données enregistrées sous forme numérique, notamment les documents sur le lieu d'entreposage, les formats et procédés utilisés, les informations techniques sur le support de données, etc. Les données de classement sont en quelque sorte les métadonnées de l'archive (et non celles du *contenu* de l'archive).
- **Métadonnées (descriptives):** Les données de l'objet sont disponibles et lisibles mais les descriptions de l'objet font défaut ou sont erronées (ex. le support de données contient des images lisibles mais les informations contextuelles de celles-ci telles que le lieu, le moment, les personnes et objets représentés, etc. sont inconnues).
- **Formats de données:** Les fichiers peuvent certes être lus sous forme d'informations binaires mais le format de données est inconnu et ne peut plus être interprété faute de logiciel adapté, par exemple.
- **Formats des supports de données:** Le support de données est illisible parce que son formatage est inconnu ou n'est plus supporté. Nous entendons par formatage d'un support de données la manière dont les données y sont enregistrées. L'existence de plusieurs formats d'écriture/lecture de données pour un type de supports de données n'est pas à exclure. Une bande magnétique du type DAT pourra par exemple être écrite aussi bien avec NT-Backup, "tar" ou en "ANSI-Labeled tape". Tous ces formats sont incompatibles: une bande magnétique écrite avec NT-Backup ne pourra par exemple pas être lue sous le format "tar". Autre exemple, les différents formats d'écriture sur les CD-R: ISO9660, Joliet, UDF, DirectCD, UDF, Mac, etc. Ces formats sont incompatibles et ne peuvent en partie être lus que sur un matériel particulier ou avec un logiciel spécial.
- **Lecteurs:** Le support de données est illisible parce que les lecteurs adaptés ne sont plus disponibles ou sont abandonnés.
- **Support de données:** Le support de données n'est plus lisible parce qu'il a été endommagé suite à l'usure, à une erreur de manipulation, etc.

Il faut donc prendre des mesures à tous ces niveaux pour pouvoir préserver l'information numérique et la garder disponible à longue échéance.

2.2. Critères de procédure

Certaines procédures, règles et stratégies doivent être respectées pour garantir la longévité des données numériques. Une archive numérique à long terme se distingue notamment par le fait que les données archivées sont utilisées en permanence. Seuls des contrôles continus et des copies/conversions permanentes pourront garantir une durée de vie (en principe) illimitée des données. Ces opérations doivent toutefois répondre à des critères de qualité très stricts et l'ensemble du processus d'archivage doit respecter certains principes.

Question: *Un concept d'archivage à long terme a-t-il été élaboré?*

- **Redondance:**

Redondance à tous les niveaux: la probabilité de perte totale des informations doit approcher de façon asymptotique la valeur zéro. La facilité de dupliquer ou de "cloner" les données numériques permet de réduire sensiblement le risque de pertes de données à tous les niveaux grâce à la redondance.

- *Redondance au niveau du support de données*

Multiplis supports de données identiques: Les données ne sont pas enregistrées qu'une fois mais de façon identique sur plusieurs supports de données. Pour une probabilité qu'un seul support de données devienne illisible de 1/1000 (une pour mille), l'existence de 3 supports de données identiques (triple redondance), réduit la probabilité de voir les trois illisibles à 1/1 000 000 000 (une sur un milliard).

Question: *Combien de copies d'un support de données sont produites et "exploitées" (contrôle qualité, migration, etc.)?*

- *Répartition géographique*

Les supports de données redondants doivent être conservés à des endroits géographiquement différents pour éliminer le risque de perte totale en raison d'une catastrophe (incendie, tremblement de terre, guerre, etc).

Question: *Les copies redondantes sont-elles conservées*

- *dans la même pièce*
 - *dans le même bâtiment*
 - *au même lieu géographique*
 - *à des endroits différents dans le pays ou dans le monde?*

- **Migration**

Une migration régulière des données étant inévitable en raison du progrès technologique ou du vieillissement des supports, il faut mettre en place une stratégie de migration à long terme. Cette stratégie doit présenter les caractéristiques suivantes:

- *Moment de la migration:*

Les données doivent être migrées suffisamment tôt pour parer à tout

risque de perte de données imputable au vieillissement du support ou à l'évolution technologique.

Question: *Les migrations sont-elles planifiées et préparées sur le plan organisationnel et technologique?*

- *Vérification périodique:*
- Les supports de données et les données qu'ils contiennent doivent être vérifiés périodiquement pour détecter par exemple un vieillissement anticipé dû à un défaut matériel.³

Question: *Les supports de données sont-ils lus périodiquement et leur consistance est-elle testée?*

- *Tolérance d'erreurs ZERO:*
Toutes les opérations de copie doivent être effectuées avec une tolérance d'erreurs ZERO pour prévenir toute perte d'information. On peut obtenir ce résultat en comparant chaque copie à "l'original" dès qu'elle a été établie.

Question: *L'exactitude de chaque opération d'écriture est-elle vérifiée (lecture de contrôle immédiatement après l'écriture des données et comparaison "original – copie")?*

- *Migration par roulement:*
L'évolution permanente des technologies de mémoire exige la migration périodique des données vers une nouvelle technologie parce que la maintenance des "vieux" appareils devient de plus en plus coûteuse, voire impossible (plus de pièces de rechange, etc.). Pour accroître la redondance par rapport à la technologie, le processus de migration d'une génération d'appareils à l'autre devrait s'effectuer par roulement: introduction d'une nouvelle technologie avant que l'ancienne ne soit obsolète. En d'autres termes, la génération "active" de supports de mémoire se répartit sur au moins deux technologies dont l'une devrait être "éprouvée".

Question: *Les migrations sont-elles prévues par roulement de sorte que les supports de l'"ancienne" génération sont encore entretenus et restent actifs pendant un certain temps après la migration?*

- **Qualité des médias**

Certains procédés se sont imposés dans la photographie analogique (pour citer un exemple) qui permettent un bon pronostic en matière de stabilité à long terme d'un certain matériel (vieillissement rapide, plan d'Arrhenius, révélation à l'acide acétique, etc.). Des procédés comparables doivent dans

³ Une telle vérification périodique peut par exemple comporter la comparaison de plusieurs supports de données (redondants) contenant des données identiques. Aucune différence ne doit être relevée à ce moment là. En cas de différences, on lira une copie supplémentaire en guise d'aide à la décision. Il en découle qu'il faut toujours établir au moins trois copies identiques.

la mesure du possible aussi être appliqués aux supports de données numériques destinés à l'archivage à long terme.

- *Lecture de contrôle*

Des erreurs d'enregistrement limitées sont inévitables avec tous les procédés d'enregistrement numériques connus à ce jour, par exemple en raison de médias imparfaits (défauts matériels). Ces erreurs d'enregistrement intrinsèques sont rattrapées en appliquant des procédés de correction⁴ ("error correction") des "recoverable errors" (erreurs récupérables). On peut ainsi garantir malgré tout une restitution correcte des données. Lorsque l'enregistreur permet de déterminer le nombre des corrections d'erreurs réussies, la qualité de la combinaison support – lecteur peut être mesurée. Par des lectures de contrôle régulières, on peut déceler une augmentation du nombre d'erreurs et on bénéficie d'un bon indicateur pour détecter rapidement des médias défectueux (vieillissant trop rapidement).

Question: *Le nombre des erreurs corrigées est-il recensé et analysé lors des lectures de contrôle périodiques?*

- *Tests des médias*

La qualité des supports de données varie d'un fabricant à l'autre et parfois aussi sur la durée chez le même fabricant. C'est la raison pour laquelle il faut vérifier la qualité par échantillons pour chaque charge de supports afin de déceler à temps les altérations de qualité dues à des défauts de fabrication (erreurs de matériel). Avec certains médias tels que les CD-R, un accord optimal entre l'enregistreur et le support est nécessaire pour obtenir une qualité optimale et donc une sécurité à long terme. Dans le cas des CD-R, on ne peut obtenir de bons résultats (CD avec très peu d'erreurs) par une combinaison graveur-disque-vitesse d'écriture optimisée. Vu que le matériel enregistreur ordinaire ne permet guère d'apprécier la qualité, il faut recourir à un hardware de contrôle spécial pour déterminer la combinaison optimale. Ce test devrait être répété pour toute nouvelle charge de médias.

Question: *Les supports de données font-ils l'objet d'un contrôle systématique pour déceler les défaillances, défauts matériels, etc. avant d'entrer dans le cycle d'archivage?*

Dans l'affirmative, cela a-t-il lieu

- *individuellement pour chaque support de données ?*
- *par échantillons pour chaque charge ?*

⁴ Ces procédés de correction d'erreurs s'appuient sur des procédures mathématiques élaborées qui permettent d'une part la détection des erreurs et d'autre part leur correction en une seule opération. Jusqu'à une certaine fréquence d'erreurs, ces procédés garantissent une reproduction absolument correcte au sens strictement mathématique des données numériques. L'un de ces procédés, largement répandu, se fonde sur le test de redondance cyclique "Cyclic Redundancy Check" et est nommé procédé CRC.

- *par échantillons pour chaque type de support de données (fabricant, fournisseur) ?*
- *généralement par échantillons sans système particulier ?*

- *Conditions de conservation optimales des supports de données (environnement, climat)*

La durée de vie des supports de données est relativement courte.⁵

Pour ne pas la réduire davantage, le stockage des supports doit répondre à certaines conditions. Pour les bandes magnétiques, on préconise une température de stockage autour de 15°C (variation de $\pm 2^\circ\text{C}$) et un degré d'hygrométrie relatif de 20 à 40% (variation de $\pm 5\%$) (SMTP, RE103 ou ANSI/AES).

Question: *Les conditions de stockage répondent-elles aux normes ou aux conditions climatiques optimales pour les différents supports de données ?*

Question: *Existe-t-il des dispositifs de protection contre les dommages causés par les éléments naturels mais aussi les cambriolages et la destruction volontaire ?*

- **Risques liés à la manipulation**

Le facteur humain représente un risque important de perte de données. On peut en effet laisser tomber un support de données, verser du café dessus ou tout bonnement l'égarer. Des erreurs de manipulation qui ont causé l'effacement d'un support de données entier se seraient déjà produites. Pour réduire au minimum les risques d'erreur humaine, il faut planifier systématiquement l'ensemble des opérations, les effectuer avec lenteur et soin et les documenter. Même une erreur minime telle qu'une erreur de marquage peut se traduire par une perte de données considérable. C'est pourquoi les opérations de traitement comportant la manipulation de supports de données doivent dans la mesure du possible être automatisées et l'ensemble du processus soumis à un contrôle qualité rigoureux.

Question: *Existe-t-il un système de contrôle qualité pour la manipulation des supports de données ?*

Question: *Les processus (archivage, lecture de contrôle, migration, etc.) sont-ils largement automatisés ?*

2.3. Critères techniques

Pour que la longévité des données numériques soit garantie, il faut remplir certains critères à tous les niveaux évoqués en vue de réduire les risques de perte de données au minimum. Examinons ces critères plus en détail:

⁵ Pour les bandes magnétiques de bonne qualité, on table sur 15 à 30 ans tandis que les CD-R peuvent atteindre dans le meilleur des cas 50 à 100 ans (estimation). C'est bien plus que la durée de vie proprement dite des mémoires de masse qui est aujourd'hui de 7 ans au maximum, compte tenu de l'évolution rapide des systèmes informatiques.

- **Données de classement / métadonnées administratives**

Les données de classement servent à organiser les supports de données et leur accès. Leur configuration la plus simple est une liste des supports de données sur lesquels se trouvent les objets à archiver. Les données de classement sont indispensables, d'une part pour retrouver les données désirées, et de l'autre, pour effectuer une migration systématique et en temps utile.

- *Les supports de données doivent être marqués de manière judicieuse mais aussi lisible pour l'homme (human and machine readable).*
Explication: Il est apparu que des erreurs se produisaient avec la meilleure organisation des opérations (erreur d'envoi postal, par exemple). Pouvoir déduire directement le contenu d'un support de données à partir de son marquage s'avère dès lors fort utile. Ainsi, une mention du type "Musée d'art de Bâle, collection Sarasin, tableaux 1-250, 12.03.1999" sera bien plus parlante qu'un numéro tel que 467812. Associée à une étiquette lisible par machine (code barres), elle peut apporter un gain d'efficacité substantiel (l'infrastructure nécessaire à l'interprétation - le lecteur de code barres - n'est pas indispensable pour autant).

Question: *Les supports de données sont-ils marqués de façon judicieuse et lisible?*

- *"Journal" des données et supports de données (version, lieu, format, date...)*
Dans le cadre du processus, il convient de tenir un journal précis de toutes les opérations de travail et plus particulièrement de tous les mouvements des supports de données. D'une part, cette démarche minimise le risque de perte et d'autre part, l'analyse approfondie de l'"historique" d'un support de données permet de déterminer les causes de problèmes éventuels (problèmes de lecture par exemple). Elle permet donc de limiter les dégâts.

Question: *Existe-t-il une journalisation ou un contrôle de qualité de ce type (selon ISO 9001, par exemple)?*

- *Données de classement également sous forme "analogique" (papier)*
Les données de classement constituent la base d'un archivage numérique à long terme sans laquelle une archive numérique se résumerait à une accumulation inutile de supports de données. Pour cette raison, les données de classement, qui sont très compactes par comparaison au contenu proprement dit de l'archive, doivent être enregistrées elles aussi avec une redondance importante. Pour permettre l'interprétation des données de classement existant généralement sous forme de texte (descriptions, listes, codes-source de logiciels) sans moyens techniques, celles-ci devraient également être disponibles sous forme analogique (tirage papier).

Question: *Comment les données de classement sont-elles archivées?*

- **Métadonnées descriptives**

Les métadonnées d'un objet (description des images, date de la prise de vue, photographe, copyright, etc.) sont certes enregistrées séparément de l'objet dans la plupart des cas mais font en définitive partie intégrante de l'objet qui serait souvent peu parlant sans métadonnées.

- *Lien métadonnées/objet*

Dans la mesure où les métadonnées font partie intégrante des données de l'objet, elles devraient, parallèlement à des méthodes d'enregistrement plus complexes telles que les bases de données relationnelles, être liées directement aux données de l'objet pour réduire au minimum le risque de perte du lien entre les métadonnées et l'objet correspondant. Il existe deux possibilités fondamentales pour les images: 1) Les métadonnées sont enregistrées dans l'en-tête du fichier⁶. 2) Les métadonnées sont directement ajoutées à l'image sous forme d'information graphique. Les métadonnées peuvent ainsi par exemple être scannées avec l'image sous forme de table de texte lors du processus de numérisation (possible même par la suite). Ceci établit également un lien direct avec l'objet à archiver. On peut souvent se servir du nom de fichier lui-même pour rendre un objet parlant.

Question: *Les ("principales") métadonnées des objets sont-elles archivées avec l'objet lui-même?*

Question: *Des noms de fichiers judicieux et parlants sont-ils utilisés?*

- *Simplicité, format horizontal*

Les métadonnées sont souvent hautement structurées et généralement organisées sous forme de bases de données relationnelles. Or, comme l'accès aux bases de données requiert des logiciels complexes qui sont également sujets à l'évolution rapide de la technologie, l'archivage à long terme des métadonnées devrait s'effectuer selon une structure simple, par exemple dans des listes ou des tableaux simples. Avec une exécution adéquate, les métadonnées archivées de la sorte peuvent à tout moment reprendre une forme structurée, nécessaire à un accès plus efficace (recherche).

Question: *Les métadonnées sont-elles archivées dans une structure horizontale?*

- *Redondance des métadonnées, lien avec l'objet*

⁶ Le format TIFF, un format d'images largement répandu, permet d'inscrire des informations complémentaires telles qu'une description, l'auteur, le copyright sous forme de texte ASCII dans l'en-tête d'une image. Les métadonnées deviennent ainsi une partie intégrante de l'image numérisée.

Les métadonnées devraient être soumises aux mêmes critères de sécurité que les données des objets elles mêmes. Il en découle qu'il faut appliquer les mêmes exigences en matière de redondance et de répartition géographique que pour les données des objets proprement dites.

Question: *Les métadonnées sont-elles archivées avec la même redondance que les données des objets?*

- *Sommes de contrôle*
Les "sommes de contrôle" peuvent également être considérées comme des métadonnées. Elles peuvent être déterminées avec des programmes appropriés (? , ?) et doivent être enregistrées avec les fichiers de données proprement dits. Les éventuelles erreurs de lecture peuvent ainsi être détectées lors de la lecture de contrôle.

Question: *Le concept d'archivage prévoit-il les sommes de contrôle?*

- **Format des fichiers**

Les formats de fichiers convenant à l'archivage à long terme doivent offrir la plus longue durée de vie possible. On peut y parvenir en satisfaisant au mieux aux critères suivants:

- *Définition ouverte*
Le format de fichier utilisé pour l'archivage doit être entièrement documenté et expliqué. Un programmeur doit en principe être en mesure de développer un programme capable de lire et d'interpréter correctement les données à partir de la documentation. C'est le seul moyen pour garantir que les données pourront être interprétées dans toutes les circonstances dans le futur (proche). Des programmes open-source⁷ qui permettent de lire et d'écrire le format existent pour de nombreux standards ouverts tels que le TIFF. Ces formats répondent idéalement à l'exigence d'une définition ouverte.
- *Diffusion*
Le format de fichier choisi devrait être le plus répandu possible. D'une part, la probabilité que le format de fichier reste longtemps d'actualité est forte. D'autre part, une large diffusion augmente aussi les chances que plusieurs prestataires commerciaux et open-source supportent ce format.
- *Flexibilité*
Le format de fichier devrait soit être évolutif, soit offrir une grande souplesse pour enregistrer conjointement des données de calibrage, des données de l'objet (par exemple la taille absolue de l'original en millimètres) et des métadonnées.
- *Tolérance d'erreurs*
Bien que la survenue d'erreurs de données soit minime avec un processus d'archivage soigneux, le format de fichier choisi devrait être

⁷ Les programmes "open-source" sont des logiciels dont le code source est public et dont la licence permet l'utilisation et des modifications/adaptations par tout le monde.

assez tolérant pour les petites erreurs (un bit erroné, par exemple). Pour les images, cela signifie par exemple qu'une petite erreur dans les données n'aurait qu'un faible impact visuel.

Selon le format, les erreurs dans les bits peuvent avoir des effets très variables. Dans l'exemple ci-dessous, l'image de droite au format JPEG comporte un bit erroné qui a un effet important (barre verte). On peut globalement partir du principe que toute forme de compression des données augmente sensiblement les effets d'une erreur dans les bits.



Avec le format d'image TIFF non comprimé, la même expérience donne une altération à peine perceptible.

Question: Quel est le format de données utilisé?

Format:	Ouverture	Diffusion	Flexibilité	Tolérance d'erreurs	Evaluation
TIFF (raw)	++	++	++	++	++
TIFF (comprimé)	++	++	++	-	-
PSD (Photoshop)	--	+	++	0	0
JPEG	++	++	++	--	-
BMP	+	+	0	+	+
PNG	++	+	++	0	+
PhotoCD	-	0	+	0	0
GIF	++	++	+	0	0
Acrobat	--	+	++	0	-

- **Format du support de données⁸**

- *Ouverture*

Il existe malheureusement de nombreux formats propriétaires pour le formatage des supports de données. Citons l'exemple des systèmes de backup. Pour l'archivage à long terme, il faut cependant choisir un format de support de données qui, soit n'est pas propriétaire, soit représente un standard industriel reconnu qui est supporté par plusieurs fabricants. Le format "tar" constitue un bon exemple de format non propriétaire: il provient initialement de l'univers Unix/Linux mais est supporté par toutes les architectures de processeurs et systèmes d'exploitation⁹. La bande NT-Backup constitue quant-à-elle certes un format propriétaire mais est largement répandue. Autres formats largement répandus: ISO9660 (CD-R), format DOS (disquettes), ANSI-labeled Tape, etc.

Question: *Quel est le formatage utilisé pour le support de données:*

- *Format standard ouvert?*
- *Format standard propriétaire? (supporté par de nombreux fabricants)*
- *Format propriétaire? (un seul fabricant)*

- **Système hardware (appareils d'écriture et de lecture)**

Le système hardware englobe le format physique du support de données (DAT, DLT, CD-R, etc.) et les systèmes d'écriture et de lecture appropriés.

- *Diffusion*

Le système de mémoire devrait être le plus répandu possible et être supporté par plusieurs fabricants. C'est d'autant plus important que même de grands fabricants peuvent rapidement disparaître du marché dans la branche informatique. La disparition du seul fabricant d'un système hardware donné remet généralement en question le support des appareils et entraîne la perte des données ainsi archivées.

- *Usure*

Le système devrait solliciter le moins possible le support de données lors de la lecture/écriture. Comme, pour tout concept d'archivage à long terme, les données doivent faire l'objet de plusieurs lectures de contrôle pendant la durée de vie d'un support de données, la sollicitation physique du support de données devrait être la plus faible

⁸ Il faudrait en fait établir encore davantage de différenciations en matière de formats: 1) Format physique du support de données, 2) Format du système de fichiers (ex. tar, FAT32, NTFS, UFS, ISO9660, etc.), 3) Format de fichier (TIFF, JPEG, etc.) et 4) Format de données (nombre de bits, ordre des bits et octets, format des chiffres à virgule flottante, etc.). Cette hiérarchie est encore compliquée par le fait que certains formats tels que tar constituent à la fois le format d'un système de données et existent eux-mêmes en tant que format de fichiers (conteneur de données). Du point de vue empirique, la répartition effectuée dans le rapport est utile et suffisante. Sur le plan du format de données, le nécessaire est généralement fait au niveau format de fichier pour que les données soient toujours interprétées correctement (TIFF, JPEG, etc.).

⁹ Si le logiciel correspondant est utilisé

possible. De ce point de vue, les supports de données optiques (disques opto-magnétiques, par exemple) peuvent donc présenter un avantage considérable. En matière de bandes magnétiques, il faut privilégier les procédés d'enregistrement linéaires au procédé "helical scan"¹⁰.

- *Fiabilité*

Le système devrait être le plus fiable possible. On la mesure d'après l'espacement des pannes dit "Mean time between failure" (MTF) qui indique la durée de fonctionnement moyenne de l'appareil sans panne. Il existe des informations MTF du fabricant pour la plupart des appareils.

- *Interchangeabilité*

Le support de données doit pouvoir être échangé sans problèmes entre différents appareils enregistreurs. L'expérience a par exemple démontré que le procédé helical-scan était peu tolérant sur ce plan. En général, il faudrait vérifier pour tout nouveau lecteur/enregistreur qu'il est interchangeable avec d'autres appareils.

- *Compatibilité des versions*

La rapide évolution technique du matériel de mémoire donne souvent naissance à des familles de technologies de mémoire pour lesquelles différentes générations du même type offrent une capacité de plus en plus importante (exemple DAT: DDS-1=2GO, DDS-2=8GO, DDS-3=24GO, DDS-4=40GO, DLT: DLT I – DLT IV). Comme la restriction provient le plus souvent de la durée de vie du système plutôt que de celle du support de mémoire proprement dit, il faut garantir une compatibilité maximale avec les générations précédentes (les nouveaux appareils peuvent encore lire les anciens supports).

- *Tolérance d'erreurs*

Le hardware devrait présenter une tolérance d'erreurs relativement importante. En d'autres termes, il doit toujours livrer des données correctes grâce à la détection et à la correction des erreurs, même en cas d'erreurs sur le support. Il doit en outre être possible de déterminer le nombre de corrections intervenues après la lecture d'un support. De nombreux fabricants indiquent à partir de combien de bits lus il faut s'attendre pour la première fois à une erreur.

- *Manipulation (robustesse, automatisation)*

Le système hardware doit être le plus robuste possible sur le plan mécanique pour réduire au minimum le risque de destruction d'un support de données par une déféctuosité ou une erreur de manipulation (!) (exemple des bandes emmêlées caractéristique des cassettes magnétiques). Enfin, le système devrait aussi permettre la plus grande automatisation possible, comme par exemple les robots de bandes pour les DLT ou DAT.

¹⁰ Le procédé "Helical-Scan", initialement développé pour les enregistrements vidéo, consiste à enrouler la bande autour d'une tête magnétique légèrement inclinée et à rotation rapide. Les données sont enregistrées sur la bande sous forme de pistes obliques.

Question: Quel est le système hardware utilisé?

Système	Conso	Usure	durée de vie	Remplacement	Compat.	Tol. Erreurs.	Robust.
CD-R	++	++	+	+	++	+	+
DVD ¹¹	+	++	+	-	0	+	+
MO	--	++	++	+	-	-	+
Disque interchangeable	+	0	0	+	-	+	0
DLT	+	++	++	++	++	+	+
Exabyte	0	-	-	0	+	0	0
DAT	+	-	-	0	++	0	+
AIT	0	+	+	0	+	0	0
LTO	0	++	++	++	++	+	+
IBM/StorageTek	+	++	++	++	++	+	+

- **Support**

- *Robustesse*

Le support lui-même devrait être le plus robuste possible pour résister sans perte de données à une manipulation brutale ou à des conditions environnementales non optimales.

- *Durée de vie*

La durée de vie moyenne du support devrait être connue et à peu près constante. Pour la plupart des supports modernes, la durée de vie physique pourrait dépasser la durée de vie du système.

- *Possibilité de vérification*

Il est possible de vérifier l'état avant l'écriture pour certains médias, par exemple par une phase d'essai consistant à y écrire un certain modèle binaire puis à le vérifier (supports plusieurs fois réinscriptibles). Pour les autres supports, on ne peut certes pas vérifier le support individuel mais le système dans son ensemble (CD-R¹²)

- *Fiabilité, dégénérescence prévisible*

Une durée de vie d'un support variant fortement et donc difficile à évaluer posera des problèmes pour l'archivage à long terme.

- *Diffusion*

Comme pour les autres critères, une large diffusion du support constitue un atout parce qu'elle garantit un support de qualité et durable du média par l'industrie.

- *Capacité*

¹¹ Tous les supports "DVD" sont regroupés sous le terme DVD: DVDRom, DVD-R, DVD+R, DVD-RW, DVD+RW, DVD-RAM

¹² Il existe des systèmes tests pour les CD-R qui permettent de vérifier la combinaison graveur-disque et vitesse d'écriture. L'expérience, notamment dans le domaine audio, a démontré que la bonne combinaison de ces trois facteurs était essentielle pour une bonne stabilité à long terme. De plus, chaque nouvelle charge de disques vierges devrait faire l'objet d'un contrôle-qualité

Plus la capacité de mémoire sera importante, moins le nombre de supports de données sera important, au même titre que le volume des données de classement. Pour une archive de 1 TB, par exemple, la manipulation de 10 bandes LTO sera plus simple et donc aussi plus sûre que la manipulation de 1500 CD-R.

Question: Quel est le support utilisé?

3. Métadonnées

3.1. Différentes formes de métadonnées

Les métadonnées sont des données au sujet d'autres données. La nature et le volume de ces données sont souvent définies de façon aléatoire, ce qui peut susciter un certain trouble. Le tableau 1 comporte les trois principales catégories de métadonnées ¹³

<i>Métadonnées descriptives</i>	<i>Description de l'objet: contenu d'une photographie, lieu, heure, etc.</i>
<i>Données de classement/métadonnées administratives:</i>	<ul style="list-style-type: none"> • <i>Information sur l'endroit et la façon dont est classé l'objet</i> • <i>Information sur l'interprétation des données numériques</i> • <i>Informations sur la numérisation (scanner, définition, etc.)</i>
<i>Métadonnées "structurelles"</i>	<i>Données sur les liens en cas d'objets complexes à plusieurs éléments.</i>

3.2. Principes pour l'archivage de métadonnées

Il y a lieu de saisir un groupe minimal de métadonnées lors de la numérisation et de l'enregistrer avec l'image numérique. La possibilité d'identifier les données sans équivoque (nom de fichier clair et judicieux) constitue le minimum absolu. La solution optimale consiste à enregistrer un petit jeu de métadonnées directement avec l'image (par exemple dans l'en-tête TIFF). L'identification sans équivoque des enregistrements est aussi indispensable pour le contrôle qualité. On peut trouver une possibilité adaptée dans le standard pour identifier des textes et images développé par la Newspaper Association of America (NAA) et l'International Press Telecommunications Council (IPTC). Ce standard comprend des mentions pour des descriptions d'objet, des mots-clés, des catégories, les droits de reproduction et l'origine. Les signatures d'images et mots clés peuvent être utilisés pour l'interrogation des banques d'images de prestataires tiers (*Adobe Photoshop®* permet par exemple la saisie selon ce standard).

3.3. Différents standards

Différents standards pour l'exploration de documents textes, étendus par la suite à d'autres médias, ont été mis au point ces dernières années. Le format de données

¹³ Le rapport "Nouvelles technologies et biens culturels (R. Gschwind, L. Rosenthaler et F. Frey, concept) nouvelles technologies et biens culturels, mai 2000, disponible à l'Office fédéral de la protection civile, section protection des biens culturels, Monbijoustrasse 91, 3003 Berne) fournit un aperçu des différents types de métadonnées.

MARC (Machine Readable Catalogue) crée par exemple un canevas permettant d'enregistrer, de lire et d'échanger électroniquement toutes les métadonnées. Le MARC est entre temps devenu un standard utilisé dans la plupart des bibliothèques.

Nous allons maintenant décrire brièvement quelques-uns des principaux standards. Vous trouverez une foule d'information sur tous les standards décrits sur Internet.

MARC (Machine Readable Catalogue)

Le MARC est un format de données pour la saisie électronique de médias. Il a été développé par la Library of Congress de Washington DC pour traiter électroniquement les informations figurant sur les fiches catalogues. Les documents sont décrits dans des champs définis numériquement. Ceci rend l'information identifiable sans équivoque et interchangeable¹⁴.

EAD (Encoded Archival Description)

L'EAD est né en 1993 à la bibliothèque de l'University of California à Berkeley. Il a été développé pour la description des matières archivées. Il applique le principe du MARC à des niveaux hiérarchiques différenciés et aux liens complexes des documents archivés. Les données sont saisies en format SGML¹⁵.

SGML (Standard Generalized Markup Language)

Le SGML a été défini en 1986 comme norme ISO 8879 pour structurer les informations électroniques. Les données sont saisies de telle manière que leur contenu et leur structure soient clairement identifiables et ne soient liées à aucune plate-forme. Un type est défini au préalable pour chaque document (DTD = Document Type Definition) et la structuration appropriée au type est établie sur cette base. Pour les pages Internet, cette DTD est le HTML.

XML Extensible Markup Language

Développé en 1997, le XML est une autre application du SGML, parallèlement à l'EAD et au HTML. Le XML offre des possibilités plus vastes que le HTML qui est progressivement supplanté par ce format. Le XML est en passe de devenir le "langage franc" de l'échange de données.

Face à ces multiples méthodes et standards, il est nécessaire de définir avec précision la forme et les directives selon lesquelles les archives seront explorées. Il est fort heureusement apparu ces derniers temps que les institutions commencent à coordonner leurs efforts afin qu'il devienne possible dans un avenir proche de rechercher et d'échanger des images non seulement à l'intérieur d'une institution mais aussi entre différentes institutions. On peut citer à cet égard l'excellent exemple du projet European Visual Archive (EVA) (<http://www.eva-eu.org>).

3.4. Métadonnées descriptives

La fonction clé de l'exploration (avec des métadonnées descriptives) consiste à placer les images dans le meilleur contexte possible. Contrairement aux

¹⁴ Vous trouverez de plus amples informations sous <http://www.loc.gov/marc>.

¹⁵ Site Web officiel: <http://www.loc.gov/ead/>.

catalogues conventionnels qui ne permettent ni la recherche, ni la présentation interactives de documents images, c'est précisément là dedans que réside l'un des atouts de l'archive électronique (voir référence 1). Le travail de préparation et de saisie manuelle des données ne doit pas être sous-estimé. Lorsque le catalogue analogique est incomplet ou n'existe pas, cette phase du projet de numérisation exige le plus de ressources en heures de travail et en finances. Il est apparu dans de nombreux projets que l'exploration doit être terminée avant d'entamer la numérisation. Le contrôle qualité est primordial dans ce domaine car une simple faute de frappe suffit à rendre une image introuvable.

L'exploration d'images est largement moins avancée que l'exploration de textes. Il y a de multiples raisons à cela. Tout d'abord, les collections d'images ont longtemps représenté un domaine spécial peu pris en considération et ensuite, l'exploration d'images soulève des questions dont la résolution exige beaucoup de temps et les compétences d'un spécialiste. (1) nota?

3.4.1. Données de classement/métadonnées administratives:

Les données de classement ou métadonnées administratives constituent une autre catégorie de métadonnées. Elles documentent toutes les procédures effectuées avec les originaux et les données numériques. Voici une liste d'exemples de métadonnées administratives:

Copyright

La situation juridique et les éventuelles conditions d'utilisation doivent être analysées et consignées.

Rapports de travail

Les rapports de travail documentent toutes les opérations. Chaque phase de travail depuis le premier scannage jusqu'à la sauvegarde des données doit être consignée pour qu'elle reste traçable (voir chapitre 3.1). Il est essentiel de standardiser les rapports de travail pour que les données restent utilisables sur différents systèmes.

Processus général

La surveillance détaillée du processus général est indispensable. Il est en effet important de disposer à tout moment d'informations sur la position d'un objet donné dans le processus. Ceci implique aussi la tenue d'un journal retraçant qui a effectué quelle tâche à quel moment.

3.4.2. Métadonnées techniques

Que sont les métadonnées techniques?

Les métadonnées techniques constituent une sous-catégorie des métadonnées administratives. Elles décrivent avec précision comment le fichier de données a été créé et permettent aussi d'enregistrer des données précises sur le système utilisé pour le scannage. Le catalogue NISO que nous allons brièvement présenter ci-dessous démontre qu'une multitude de données peuvent être enregistrées au titre de métadonnées administratives. Il faut toutefois se rendre compte que seul

un certain nombre d'éléments est requis. Une grande partie des éléments peut mais ne doit pas forcément être enregistrée.

Catalogue de métadonnées NISO

NISO a commencé depuis deux ans à élaborer un catalogue de métadonnées pour les images numériques (www.niso.org). L'objectif était d'établir un catalogue de métadonnées techniques nécessaires à la gestion de fichiers d'images numériques. Ce catalogue ("Technical Metadata for Digital Still Images") ne s'adresse pas qu'aux institutions culturelles mais aussi à toutes les personnes qui créent à titre professionnel des fichiers de données d'images à partir d'objets appartenant à des collections.

Le catalogue de données n'est pas tributaire d'un format de fichier particulier. Le catalogue de métadonnées NISO admet aussi d'autres standards courants, ce qui facilitera son application dans les domaines les plus divers. La première phase de ce catalogue de données est terminée. Dans la prochaine phase, il sera testé en version bêta par différentes institutions. Cette phase devrait se terminer au plus tard vers le milieu de l'année 2003. La description des éléments permet de réaliser très simplement un catalogue DTD. Dans la plupart des cas, les métadonnées NISO seront codées en XML. Voici quelques catégories de données:

- *Basic Image Parameters (paramètres de base image)*
Ces éléments permettent la visualisation du fichier numérique sur un écran.
- *Image Creation*
Cette section regroupe toutes les indications sur la numérisation. Il est également important dans ce contexte de consigner le nom de la personne ayant commandé le scanner. Cette section comporte de nombreuses données enregistrées automatiquement par le scanner lors du scannage, entre autres l'heure exacte du scannage. Une fois enregistrée, cette indication ne devra plus jamais être modifiée.
- *Image Performance Assessment (évaluation des performances de l'image)*
Les données dans cette section devraient garantir la qualité des fichiers numériques. D'une part, des modèles-test doivent servir à surveiller la qualité des scannages au moment de la saisie et d'autre part, ils doivent permettre de transmettre les scans à un autre système dans le futur sans perte de qualité. Il s'agit en particulier d'appuyer la migration comme stratégie d'archivage. Les modèles-test répondent aux modèles décrits sous la rubrique "Contrôle qualité" (chapitre ?)
- *Change History (historique des modifications)*
Cette section sert à consigner tous les processus (édition ou transformation) subis par les images. Ce bloc de données n'a pas vocation à annuler les processus mais à permettre l'évaluation de la qualité des données.

Création des métadonnées techniques

Les métadonnées techniques sont créées à différentes étapes du processus. Un grand nombre d'entre elles sont automatiquement incorporées à l'en-tête du fichier lors du scannage (voir tableau ?). Les autres données sont ajoutées automatiquement ou manuellement lors du contrôle. Il est essentiel de contrôler ces métadonnées pour garantir qu'elles seront conformes au contenu du fichier.

Métadonnées minimales

Un jeu minimal de métadonnées devrait être constitué à chaque scannage. Il faut toutefois garder à l'esprit que ce jeu minimal ne suffit pas comme métadonnées de conservation (preservation metadata).

Dublin Core

Le Dublin Core a été créée en 1995 par l'OCLC (Online Computer Library Center) à Dublin, Ohio. Il a initialement été développé pour structurer les données électroniques sur Internet. Il s'applique toutefois aussi aux données physiques et convient pour toutes les matières visuelles. Dans sa forme actuelle, le Dublin Core est composé de 15 éléments qui suffisent pour la saisie des informations essentielles, raison pour laquelle il ne peut être comparé aux règles complexes classiques. Le Dublin Core est depuis quelques semaines un standard international. Vous trouverez des informations détaillées sous www.dublincore.org.

Les 15 éléments du Dublin Core (état février 2002)

Title	Titre du document
Creator	Nom de l'auteur
Subject	Restitution standardisée du sujet
Description	Description du contenu
Publisher	Nom de l'éditeur
Contributor	Nom de la personne qui a contribué à la réalisation
Date	Date liée au document
Type	Attribution à un genre
Format	Caractéristiques spécifiques au document
Identifiant	Identification sans équivoque du document
Source	Source ayant servi de modèle au document
Language	Langue du contenu intellectuel
Relation	Lien avec un document parent utilisé
Coverage son contenu	Portée historique ou géographique du document ou, en l'occurrence de
Rights	Informations sur la situation juridique du document

Chacun de ces 15 éléments est de son côté caractérisé par dix attributs. On y trouve notamment le schéma type selon lequel les informations sont inscrites dans l'élément. Les attributs précisent par ailleurs si un élément est obligatoire, s'il peut être répété et qui a défini ces dispositions.

Il ne faut toutefois en aucun cas perdre de vue que le Dublin Core n'est en réalité qu'un jeu minimal de métadonnées. Dans certains projets, les institutions ont décidé de créer ce jeu de métadonnées minimal de façon uniforme pour permettre ainsi les échanges et recherches sur plusieurs collections. Il existe cependant souvent de nombreuses métadonnées au plan interne à partir desquelles le Dublin Core est "mappé". Le projet EVA illustre une fois encore parfaitement cette démarche.

Schéma de codage

Il existe différentes possibilités pour l'endroit et la façon dont les métadonnées peuvent être codées. Trois d'entre elles sont aujourd'hui appliquées dans la plupart des cas:

- Les métadonnées sont insérées dans l'en-tête des fichiers de données. Le volume de données incorporable dépend alors du format de film utilisé. Il est évident qu'il faut définir avec précision quelles données se trouvent dans quel champ de l'en-tête. Cette méthode ne convient toutefois pas pour une interrogation rapide des métadonnées parce que le fichier doit être ouvert.
- Les métadonnées sont enregistrées dans une base de données séparée. Celle-ci comportera les métadonnées enregistrées automatiquement dans l'en-tête des fichiers mais aussi les données saisies manuellement. Il est primordial de garantir le lien entre le fichier de données et l'enregistrement de données, par exemple pour que les modifications soient consignées de façon consistante dans le fichier de données.
- Les métadonnées sont déduites "au vol" par exemple d'un enregistrement MARC.

Il faut en tout cas assurer que les métadonnées soient toujours à jour et que les données dans l'en-tête concordent avec celles dans la base de données en cas de modifications.

On doit décider comment les métadonnées seront codées, quelle que soit la méthode choisie. De plus en plus d'entreprises se mettent ainsi à coder les données en XML. L'avenir nous dira comment ce mouvement évoluera. Il est en tout cas clair que seul un schéma de codage standardisé permettra un échange de données simple.

Où les métadonnées sont-elles insérées dans le cycle de travail?

Les métadonnées sont insérées aux différentes étapes du processus général. Il conviendrait de constituer un jeu entier de métadonnées *descriptives* avant d'entamer la numérisation. A défaut, on risque que la numérisation avance plus vite que la création des métadonnées descriptives, ce qui peut se traduire très rapidement par un retard.

Une autre partie des métadonnées est automatiquement insérée lors de l'opération de scannage. Les autres métadonnées doivent être saisies manuellement aux différentes étapes du processus. Il est important que la saisie et l'opération de scannage soient rapprochées dans le temps car c'est le seul moyen de garantir que les éventuelles erreurs puissent être immédiatement corrigées. Un contrôle qualité rigoureux des métadonnées est lui aussi primordial. C'est le seul moyen de garantir que toutes les informations nécessaires pour l'archivage à long terme seront disponibles sous la forme appropriée. En conclusion, le paragraphe suivant se propose encore d'expliquer quelles métadonnées sont nécessaires pour un archivage à long terme.

Métadonnées pour l'archivage à long terme

Les Preservation Metadata, métadonnées pour l'archivage à long terme, sont constituées pour l'essentiel de métadonnées administratives et structurelles. Les aspects suivants sont à prendre en compte:

- Les données techniques servant de base aux décisions pour l'archivage à long terme doivent être enregistrées.
- Les procédures en vue de l'archivage à long terme telles que la migration ou l'émulation doivent être documentées avec précision.

- Les répercussions des procédés choisis pour l'archivage à long terme doivent être consignées.
- Il faut veiller à ce que l'authenticité des données numériques puisse être durablement garantie.
- Les informations en matière de copyright doivent être consignées avec précision et être à jour.

Le Research Libraries Group (RLG) a publié le jeu de métadonnées suivant pour l'archivage à long terme en 1998:

Date
Transcriber
Producer
Capture Device
Capture Details
Change History
Validation key
Encryption
Watermark
Resolution
Compression
Source
Color
Color management
Color bar/grayscale bar
Control targets

Il ressort de la liste qu'il s'agit de métadonnées administratives et structurelles. Cette liste ne doit pas être considérée comme un standard mais servir de base. Il sera important de suivre des projets portant sur l'archivage à long terme de données numériques menés à grande échelle.

Les projets suivants doivent servir d'exemples pour la méthode d'archivage à long terme et les stratégies à appliquer en matière de métadonnées (Les adresses Internet indiquées tiennent lieu de référence pour des informations détaillées sur les différents projets).

- CEDARS (CURL Exemplars in Digital Archives Project)
- National Library of Australia
- NEDLIB (Networked European Deposit Library)
- Harvard University's Digital Repository Services (DRS)

Les trois premiers projets suivent le modèle OAIS (Open Archival Information System Reference Model) (www.ccsds.org/documents/pdf/CCSDS-650.0-R-1.pdf).

Le modèle de référence OAIS constitue une trame conceptuelle pour une archive numérique. Il établit une terminologie et des concepts déterminants pour l'archivage numérique, identifie les composants clés et procède de façon caractéristique pour la plupart des activités d'archivage numérique. Il propose un modèle d'information sur les objets numériques et les métadonnées associées. Ce modèle de référence ne spécifie pas de procédure de mise en œuvre et est de ce fait neutre par rapport aux types d'objets numériques et aux aspects techniques (traduction de l'anglais).

Harvard University's Digital Repository Services (DRS) utilise une structure XML pour les métadonnées de conservation.

Indépendamment de la méthode choisie, ces quatre projets démontrent que les métadonnées de conservation doivent être indépendantes du type de l'objet numérique et de la technologie utilisée pour l'archivage à long terme. C'est la raison pour laquelle on peut appliquer une seule méthode uniforme pour les métadonnées de conservation à un grand nombre des différentes activités d'archivage à long terme.

3.5. Références

- 1) Kathryn Pfenninger, Bildarchiv Digital, Sonderheft 7 Rundbrief Fotografie, 2001.
- 2) Cedars project www.leeds.ac.uk/cedars/MD-STR~5.pdf
- 3) Consultative Committee on Space Data Systems (1999) Reference Model for an Open Archival Information System www.ccsds.org/documents/pdf/CCSDS-650.0-R-1.pdf
- 4) Harvard University Library (2000) Digital Repository Services (DRS): User Manual for Data Loading: A Guide for Producers of Digital Still Images (version 1.9)
- 5) National Library of Australia (1999) "Preservation Metadata for Digital Collections: Exposure Draft" www.nla.gov.au/preserve/pmeta.html
- 6) Networked European Deposit Library (2000) "Metadata for Long Term Preservation" <http://www.kb.nl/coop/nedlib/results/preservationmetadata.pdf>

4. Directives pour la numérisation de photographies

4.1. Introduction

De nombreux musées et archives planifient ou réalisent déjà des projets de numérisation de collections photographiques. Parmi les multiples raisons, on peut citer la préservation des biens culturels, la pérennisation de la photographie en alternative à la micromation photographique. Pourtant, bien souvent un projet de numérisation ne s'inscrit pas dans la perspective d'un archivage à long terme mais par exemple dans celle d'une amélioration de l'accès à la collection, d'un site Internet, etc. Les exigences d'un projet de numérisation, la gestion du projet, le processus général peuvent s'avérer complexes et requièrent une bonne préparation. Vous trouverez un processus schématique en annexe.

Lorsque l'archivage à long terme est l'un des objectifs de la numérisation, il faut tenir compte d'exigences et de critères de qualité bien particuliers¹⁶. Dans ce contexte, nous partons du principe que l'image numérique doit servir de substitut à l'original pour le continu visuel, en d'autres termes que l'image numérique puisse pleinement remplacer l'original lorsque seul le contenu visuel est déterminant. Le critère général pour l'évaluation de la qualité de numérisation en découle

Le "fac-similé numérique" contient au moins autant d'informations visuelles que celles qui peuvent être extraites de l'image originale avec des moyens conventionnels dans le domaine des techniques de reproduction.

naturellement:

Cependant, une numérisation d'une qualité suffisante ne suffit pas à satisfaire toutes les exigences d'un processus de numérisation en vue d'un archivage à long terme. Un tel processus doit en effet répondre à des exigences de qualité spéciales sur les trois plans suivants:

- *Numérisation*

La numérisation proprement dite (ou le scannage) sert de trait d'union entre l'univers analogique (photographie classique) et l'univers numérique de la reproduction informatisée (code numérique au traitement automatisé). Ce processus de traduction doit être suffisamment fidèle pour restituer le contenu visuel de l'image en code numérique. Enfin, on définit aussi ici la limite supérieure de la qualité de l'image numérique. A l'exception d'une nouvelle numérisation, il n'existe aucun procédé permettant d'améliorer par la suite une numérisation insuffisante.

- *Métadonnées*

Les métadonnées (informations sur l'image) doivent être saisies pour chaque image et être classées de façon judicieuse. Une image sans les métadonnées qui s'y rapportent (moment, lieu de la photographie,

¹⁶ Il faut toujours garder à l'esprit que la numérisation en vue de l'archivage à long terme n'est pas un acte conservatoire (conservation de la photographie originale) mais correspond plutôt à la micromation.

circonstances, descriptions, etc.) ne présente aucune valeur sur le plan de la teneur culturelle.

- *Contrôle qualité*

La numérisation doit aller de pair avec un contrôle qualité permanent. Des erreurs à même de rendre les images numériques inutilisables pour un archivage à long terme peuvent survenir dans la numérisation proprement dite comme dans les métadonnées, en raison de défaillances techniques mais aussi d'erreurs humaines.

4.2. Numérisation

La numérisation doit répondre à divers paramètres parce qu'il faut par principe viser la plus haute qualité possible. L'objectif est de transposer intégralement l'information visuelle de la photographie originale dans le numérique, c'est à dire produire un maître numérique ou le fac-similé numérique.

Lors de la numérisation, l'"image photographique" est décomposée en unités discrètes, un nombre déterminé de points (généralement carrés) ou de pixels (définition spatiale). La "luminosité" de chaque pixel est également décomposée en unités discrètes lors de la numérisation (définition photométrique). La nature de la déperdition d'informations par rapport à l'image numérique est donc double:

- Les scanners utilisés pour la numérisation font appel à un système optique. En d'autres termes, l'image (encore analogique) du modèle photographique souffre forcément de déperditions d'informations (erreurs de représentation, diffraction de la lumière, erreurs chromatiques, lumière parasite)¹⁷.
- L'information existant dans l'image analogique subit encore des déperditions supplémentaires en raison du tramage et de la discrétisation consécutive.

Il faut néanmoins intégrer que l'information visuelle de la photographie originale est elle aussi déterminée. En effet, l'appareil photo, la pellicule, l'agrandisseur et le papier photo ont une qualité de reproduction limitée et donc une définition limitée. Une qualité déterminée de la numérisation qui s'adapte à la qualité du modèle photographique est donc suffisante. A part un surcroît inutile du volume de données, une qualité de numérisation supérieure n'apporte aucun avantage et doit par conséquent être évitée.

Les critères décrits ci-après s'appliquent au processus de numérisation proprement dit:

4.2.1. Choix du matériel

Il faut tenir compte des deux aspects suivants lors du choix du matériel de numérisation:

- Qualité de reproduction analogique/électronique du scanner
Cette qualité doit être équivalente sur le plan analogique (optique, mécanique) et sur le plan numérique (électronique, CCD, conversion AD).
Vous trouverez plus de détails à ce sujet au point "contrôle qualité".

¹⁷ Cette déperdition d'information se présente aussi avec la micromatation.

- Adaptation physique aux originaux
L'appareil doit être adapté à l'original photographique (souvent délicat) pour éviter qu'il soit endommagé physiquement lors de la manipulation. Ces critères doivent à chaque fois être clarifiés individuellement. C'est ainsi que les scanners à plat sont généralement inadaptés pour les albums photo, par exemple. Certains matériels ne doivent pas être pressés sur le verre, l'échauffement dû à l'éclairage ne doit pas être trop fort¹⁸, les négatifs en verre ne doivent pas être posés sur la vitre parce qu'ils risquent d'y rester collés et pourraient se briser lorsqu'on les retire du scanner...

4.2.2. Définition spatiale

La définition spatiale de l'image numérique, en l'occurrence le nombre de pixels (largeur fois hauteur), est fonction du modèle photographique, à savoir de sa "teneur en information". Celle-ci est constituée de la définition des matériels photographiques et de la capacité de reproduction du système optique. Dans la photographie, le critère de mesure de ce paramètre est la fonction de transfert de modulation (FTM). Il est difficile d'obtenir des chiffres exacts à ce propos dans la pratique. S'il existe des mesures pour les matériels modernes, il faut se baser sur des valeurs empiriques avec le matériel historique (dont la qualité de reproduction est clairement inférieure).

Lors du scannage, la définition est mesurée en dpi (dots per inch – points par pouce) et les définitions minimales suivantes doivent être respectées:

Diapositives et négatifs 35mm – 6x6	= 2700 dpi=
6x9 – 4/5"	= 2000 dpi
13x18 – 8/10"	= 1200 - 1500 dpi
Négatifs en verre historiques jusqu'à 13x18	= 1200 - 1500 dpi
Négatifs en verre historiques 18/24	= 900 dpi
Tirages papier n/b ou couleur	= 500 dpi

4.2.3. Valeur tonale

Les "tons" d'une photographie (restitution de la luminosité) doivent être numérisés intégralement et objectivement. Dans le numérique, c'est le nombre de niveaux de gris ou de couleurs disponibles pour la représentation de la valeur tonale qui est déterminant. Ce paramètre est mesuré en bits (par valeur mesurée) 8 bits permettent de représenter $2^8 = 256$, 12 bits $2^{12} = 4096$ et 16 bits $2^{16} = 65536$ niveaux de valeur tonale. La profondeur du bit est elle-même déterminée par la nature de l'original et son contraste. Nous devons naturellement définir les notions de valeur tonale, de valeur des gris ou de niveaux de couleur. S'agit-il de références physiques telles que la transmission/réflexion (unités linéaires) ou de densités optiques (unités logarithmiques)? S'agit-il d'une perception de luminosité

¹⁸ L'influence photochimique de la lumière ne joue en soi aucun rôle mais l'original ou l'appareil peuvent s'échauffer en cas d'utilisation prolongée. Or, cet effet thermique doit être pris en compte.

ou de valeurs de "correction gamma" (relativement peu normalisées) pour rendre linéaires les lignes de référence du moniteur.

L'observateur humain est capable de différencier simultanément un maximum de 100 niveaux de luminosité lors de l'observation globale d'une image tandis que l'amplitude des brillances se mesure à une échelle de 1 à 100. Tout ce qui est plus sombre est perçu comme étant "noir". De ce point de vue là, 8 bits par unité de mesure pourraient a priori suffire. Or la vue humaine a la faculté de s'adapter aux conditions globales. En d'autres termes, pour une image à dominante claire, la centaine de niveaux ne sera répartie que sur la partie claire et, par analogie, sur la partie sombre pour les images à dominante sombre. De plus, on peut augmenter localement le contraste par des méthodes photographiques (agrandissement de détails, etc.). Par ailleurs, la perception humaine de la luminosité suit approximativement une échelle logarithmique tandis que les capteurs de lumière électroniques (CCD) suivent une échelle linéaire. Le fac-similé numérique devra tenir compte de toutes ces possibilités, raison pour laquelle 8 bits par valeur mesurée offrent généralement une définition insuffisante.

La valeur tonale d'un matériel photographique est déterminée d'après la courbe "caractéristique" Deux éléments sont importants pour la numérisation: le contraste, à savoir le rapport entre le point le plus clair (D_{\min}) et le plus foncé (D_{\max}) et le degré d'exposition, à savoir le rapport entre le point le plus clair et le plus sombre au niveau du motif (ou de la scène) qui peut être enregistré sur le matériel photographique¹⁹

Matériel photo	Contraste maxi	Degré d'exposition maxi	
Diapositives	1:1000	1:100	
Négatif couleur négatif n/b "standard"	1:50	1:1000	
Négatifs en verre historiques		1:10'000	1:1000
Agrandiss. papier	1:100	-	

Le tableau ci-dessus démontre que ces données sont très variables. La numérisation des matériels négatifs est particulièrement problématique car les parties importantes de l'image sur le plan visuel sont sombres sur le négatif et seraient enregistrées sans différenciation par le scanner avec une définition des niveaux de gris trop faible.

On peut donc prendre en compte des exigences minimales mais des investigations complémentaires sont nécessaires en fonction du modèle:

Originaux transparents positifs (diapo)	= 12 bits
Modèles transparents négatifs	= 14 bits
Original positif (tirage) linéaire	= 10 bits

¹⁹ Le degré d'exposition associé à la définition (netteté) donnent la mesure de la teneur en information d'un matériel photographique.

Original positif "logarithmique"	= 8 bits
----------------------------------	----------

Ce tableau ne reflète que des valeurs indicatives. Il faut noter que ces chiffres se réfèrent à des bits effectivement mesurés. Vous trouverez des indications sur la valeur en bits effectivement mesurée dans les informations techniques du fabricant du scanner. Elles devraient aussi être vérifiées par des mesures régulières (voir contrôle qualité).

Pour des raisons pratiques (architecture informatique), les valeurs de luminosité ne sont possibles qu'en format 8 bits ou 16 bits lors de l'enregistrement des données (pour la couleur par analogie 3 x 8 bits = 24 bits ou 3 x 16 bits = 48 bits). Il convient donc d'appliquer les formats suivants pour un archivage à long terme:

Négatif	16 bits ²⁰
Diapositives	16 bits ²¹
Originaux positifs / tirages	8 bits

Pour les originaux positifs il est possible représenter les valeurs de 10 – 12 bits existants au plan interne, par exemple sous forme logarithmique.

4.2.4. Calibrage et reproduction des couleurs

Lors de l'archivage, on doit connaître la corrélation entre la valeur en pixels pour la luminosité et la couleur. Le scanner doit donc être calibré (v. également contrôle qualité). Les modèles suivants sont disponibles comme étalons de calibrage:

Pour les modèles n/b:

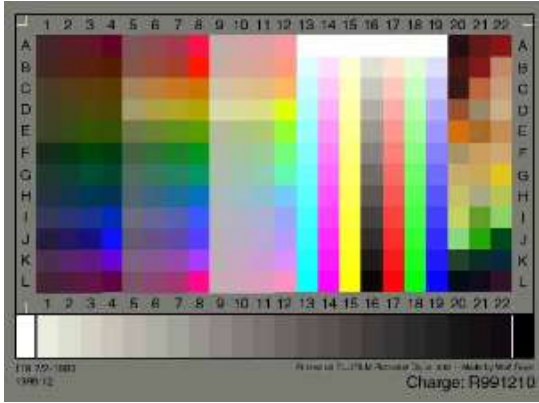
coins photométriques (positifs et transparents) que l'on peut obtenir auprès de l'industrie photographique et qui peuvent être mesurés de façon relativement simple avec un "densitomètre".

Pour les modèles en couleurs:

a) plusieurs fabricants proposent des "étalons de calibrage de scanners (scanner calibration targets" au standard IT 8.7 et ce, à la fois pour les films diapos et le papier photo.

²⁰ Les négatifs doivent être archivés en qualité de "négatifs" et non être convertis en positifs au niveau de la valeur tonale.

²¹ En présence de diapos couleur sans grandes exigences par rapport à la teneur en informations (documentations de masse, p. ex.), un enregistrement en 3x8 bits se justifie.



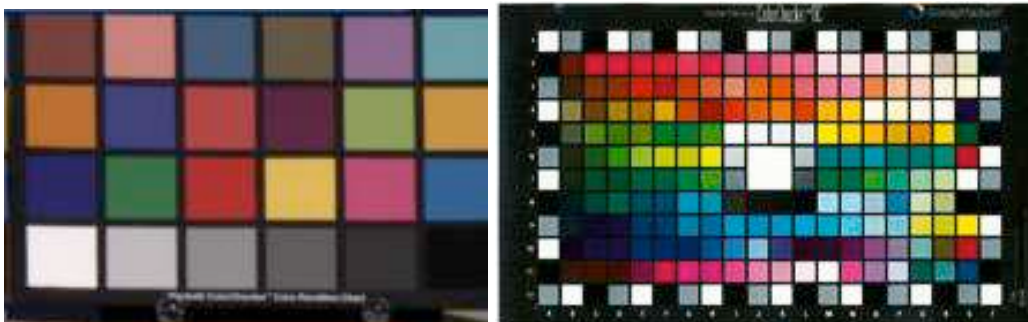
Ces étalons sont réalisés sur du matériel photo couleur et sont livrés calibrés. La valeur chromatique correspondante est attribuée à chaque zone de couleur, par exemple (extraits):

```

.
.
SERIAL "5x7 R991210"
MATERIAL "FujiFilm Pictrostat"
NUMBER_OF_FIELDS 9
BEGIN_DATA_FORMAT
SAMPLE_ID XYZ_X XYZ_Y XYZ_Z LAB_L LAB_A LAB_B LAB_C LAB_H
END_DATA_FORMAT
NUMBER_OF_SETS 288
BEGIN_DATA
A1 3.39 2.73 1.85 18.91 13.35 3.83 13.89 16.01
A2 4.58 2.98 1.63 19.97 26.04 7.92 27.22 16.92
A3 5.66 3.32 1.47 21.28 33.57 12.06 35.67 19.77
A4 5.93 3.19 1.49 20.78 38.78 10.98 40.30 15.81
A5 13.33 11.56 8.14 40.51 14.96 5.02 15.77 18.54
A6 15.27 11.21 6.98 39.93 29.46 8.65 30.70 16.37
A7 17.05 10.77 5.96 39.20 42.75 11.86 44.36 15.50
A8 21.82 10.52 4.43 38.76 68.64 18.94 71.21 15.43
A9 41.15 40.44 31.91 69.78 6.68 2.20 7.03 18.22
.
.

```

b) On utilise souvent le Macbeth ColorCheker comme étalon du positif et le Macbeth ColorChacker DC (Digital Camera) pour les appareils photos numériques²².



²² Le nouveau nuancier "Macbeth Colorchecker DC" avec 117 champs de couleur a été spécialement développé pour la photographie numérique et les appareils photo numériques.

Ces étalons devraient également être scannés lors de la numérisation et archivés. Il est important dans ce contexte a) que cela soit effectué régulièrement et b) que la numérisation soit effectuée avec des paramètres de scannage constants.

Il subsiste un problème non résolu pour le matériel négatif photo pour lequel il n'existe rien (et n'existera certainement jamais rien) d'analogique en matière d'étalons IT8. Les négatifs couleurs ne comportent pas de "couleurs" au sens visuel: celles-ci ne sont produites qu'au travers d'un processus de copie photographique (par exemple agrandissement sur papier) mais il n'existe pas de couleur "objective et incontestable" ("le laboratoire détermine la couleur"). Notre seule recommandation est d'enregistrer les négatifs couleurs en RGB 3x16 bits. La conversion en positif demeure toutefois une démarche subjective.

Les méthodes de gestion des couleurs se fondent sur le fait que les étalons numérisés (IT8.7, voir plus haut) peuvent être comparés avec les valeurs chromatiques actuelles. Des tables de conversion en sont déduites (par approximation et interpolation). Ces profils permettent de contourner les propriétés spécifiques au scanner et on obtient (plus ou moins) d'informations sur les couleurs qui sont indépendantes du scanner (device independant). Les méthodes de gestion de la couleur ne conviennent pas (encore) pour l'archivage à long terme parce que le développement de ces techniques n'est pas encore terminé et que les "standards" évoluent en permanence.

Nous recommandons pour le calibrage des valeurs de gris et la restitution des couleurs:

- La numérisation régulières de modèles-étalons (coins photométriques, IT8, Macbeth ColorChecker ou Macbeth ColorChecker DC)
- Numérisation avec des paramètres de scannage constants

4.2.5. Courant d'obscurité, bruit (bruit) et sensibilité

Les capteurs photosensibles électroniques tels que les CCD sont, comme tout système électronique, soumis à différentes perturbations dont l'effet peut être réduit par une correction appropriée.

- *Courant d'obscurité*

Un capteur photosensible électronique émet un petit signal, même dans l'obscurité absolue, qu'on appelle courant d'obscurité. Ce signal parasite est différent pour chaque élément d'une grille CCD et doit être mesuré à l'aide d'une prise de vue sur fond noir. Cette opération consiste à numériser une image sans éclairage à partir d'un modèle opaque ou noir.

- *Bruit (électronique)*

Le courant d'obscurité varie de façon aléatoire dans certaines limites. En d'autres termes, il n'est pas forcément identique lors de deux mesures consécutives sur un élément CCD donné (bruit). Ce caractère variable ne doit toutefois pas dépasser la précision théorique du scanner. En

conséquence, seul le dernier bit peut varier de façon aléatoire avec une définition à 12 bits garantie. Avec une électronique "non optimale" on peut aussi rencontrer en plus des structures périodiques nommées "fixed pattern noise" qui sont très perturbatrices sur le plan visuel. Le bruit doit être vérifié périodiquement dans le cadre du contrôle qualité parce qu'il peut se modifier sensiblement avec le vieillissement de l'électronique.

- *Bruit (bruit quantique, bruit lumineux)*

Avec une faible exposition, chaque élément photosensible d'un CCD n'enregistre que peu de photons. Un bruit (physiquement inévitable) se produit même avec un capteur photoélectrique "idéal" sans bruit d'obscurité. Il est proportionnel à la racine du nombre de quantas lumineux entrants, ce qui donne la "statistique photonique". Ce type de bruits peut poser des problèmes avant tout lors de la numérisation de négatifs noirs et surexposés.

- *Sensibilité*

Pour un même volume de lumière, les différents éléments d'une grille CCD ne donnent pas forcément le même signal. C'est pourquoi il faut effectuer une mesure en champ clair qui consiste à mesurer le signal de chaque élément CCD en situation d'exposition maximale (p. ex. sans modèle ou avec modèle transparent). Ceci permet de prendre aussi en compte les effets locaux de l'éclairage.

Les données brutes sont alors corrigées (pour chaque point de l'image) selon la formule suivante:

$$I(x,y)_{cal} = \frac{(I(x,y)_{mess} \square I(x,y)_{dunkel}) \cdot m}{I(x,y)_{hell} \square I(x,y)_{dunkel}}$$

m constituant la valeur maximale possible (pour les données 8 bits, par exemple $m=255$) et I la valeur mesurée non corrigée.

4.2.6. Distorsions géométriques

L'optique des scanners peut causer de légères distorsions géométriques. Celles-ci sont en général négligeables mais peuvent aussi requérir une correction géométrique pour des modèles spéciaux (clichés métriques tels que des photos aériennes, des images pour l'enregistrement d'objets d'art en 3D, etc.). Les distorsions géométriques peuvent être facilement décelées en scannant par exemple du papier millimétré.

4.3. Contrôle qualité

Un contrôle qualité strict à tous les niveaux est indispensable lors des numérisations qui ont entre autres pour objectif l'archivage à long terme.

- **Hardware**

Le matériel utilisé, et plus particulièrement le scanner, doit être périodiquement vérifié et étalonné. Les paramètres déterminants devront être documentés et comparés avec les mesures antérieures. En effet, les altérations significatives

des paramètres indiquent fortement un problème d'ordre matériel (vieillesse de l'éclairage, vieillissement du module CCD, alimentation électrique insuffisante et variations de la tension, etc.)

- Si le scanner n'y procède pas automatiquement, il faut corriger le courant d'obscurité et la sensibilité (correction du "shading") au moins une fois par semaine, l'idéal étant de le faire quotidiennement.
 - L'étalonnage devrait également être effectué chaque semaine.
 - Le comportement en matière de bruit devrait être mesuré mensuellement.
- **Contrôle visuel des images numérisées**

Toutes les images numérisées de haute définition doivent être soumises à un contrôle visuel. On portera une attention particulière aux caractéristiques suivantes:

- *Netteté*
L'usure mécanique, les désalignements, les chocs, etc. peuvent altérer la géométrie de l'optique et provoquer ainsi un flou systématique.
- *Poussière et saletés*
La poussière, des originaux sales, etc. peuvent causer des dépôts sur la voie optique du scanner (par exemple la vitre d'un scanner à plat) suscitant des parasites intolérables dans l'image numérique. Les vieux originaux notamment peuvent souvent laisser des traces sur le scanner qui peuvent apparaître sous forme parasite sur les images suivantes et exiger d'importantes retouches.
- *Géométrie*
Les images sont-elles numérisées selon une géométrie correcte et non inversées (en miroir)?
- *Erreurs de scannage*
L'apparence de l'image est-elle conforme aux attentes (piqué des couleurs, luminosité, intégrité, etc.)?

Le contrôle visuel peut intervenir directement lors du processus de numérisation. Pour des questions d'efficacité, un contrôle visuel approfondi pendant la numérisation est cependant souvent difficile. Celui-ci doit néanmoins être fréquent (au moins quotidien) pour déceler au plus tôt les erreurs systématiques telles que des saletés sur la vitre du scanner. Il y a alors lieu de rescanner les images concernées.

- **Intégrité**
La numérisation de stocks d'images en vue d'un archivage à long terme comprend souvent un nombre important d'images formant un ensemble. Or, avec la fatigue, l'inattention ou par d'autres facteurs, on peut facilement "oublier" de scanner certaines images, attribuer des noms de fichiers deux fois (ou des noms erronés en raison d'erreurs de frappe) de sorte que la première image soit écrasée, etc. Lorsqu'on a scanné un ensemble, il faut impérativement en vérifier l'intégrité:

- Le nombre d'images numériques est-il conforme au nombre d'originaux?
- Les noms de fichiers sont-ils conformes aux attentes?
- Les métadonnées attribuées à l'image sont-elles consistantes avec son contenu?

Ces vérifications devraient être documentées.

Un tel contrôle qualité garantit une base de données adaptée à un archivage à long terme. Tandis que certains éléments du contrôle qualité ne peuvent être exécutés que par un contrôle humain, d'autres peuvent être largement automatisés. Nous recommandons une automatisation la plus large possible pour les grands stocks d'images alors qu'un contrôle par intervention humaine suffit pour les petits stocks.

5. Annexe

5.1. Exemple d'archivage à long terme

A l'institut des sciences des médias, section technologies de l'image et des médias (anciennement section de la photographie scientifique) de l'université de Bâle nous effectuons déjà depuis un certain nombre d'années non seulement des numérisations, mais aussi un archivage à long terme pour différents musées et archives. Nous souhaitons démontrer ici comment nous procédons. L'archivage à long terme effectué à l'institut garantit une très haute sécurité des données même s'il n'est pas toujours possible de choisir le procédé optimal pour diverses raisons (financement, manque de place, etc.).

5.1.1. Numérisation et contrôle qualité

Les paramètres de numérisation (définition, etc.) sont déterminés en étroite collaboration avec le donneur d'ordre pour atteindre une qualité optimale des images. Pour l'archivage, nous enregistrons les images en format TIFF non comprimé. Ce faisant, nous exploitons largement les possibilités offertes par le format TIFF pour enregistrer un petit jeu de métadonnées dans l'en-tête du TIFF (calibrage, par exemple). L'archivage proprement dit se fait sur des bandes magnétiques DLT²³ ou LTO-ultrium à l'aide du programme `tar` (**t**ape **a**rchiver). Le `tar` est entièrement documenté, open source et il en existe des implémentations pour tous les systèmes informatiques courants. La préparation, l'archivage et la migration des données images sont effectuées sur des systèmes Linux. Le déroulement est le suivant:

1. Les images sont dégagées et si nécessaire redressées (tournées). Ce premier pas se fait de façon partiellement interactive et sert en même temps à un premier contrôle visuel des données numérisées. Pour effectuer cette opération de façon sûre et efficace, nous avons développé des programmes spéciaux à cet effet, qui se fondent sur un mélange de traitement interactif et de traitement automatisé par piles.
2. Les données de calibrage sont saisies et inscrites dans l'en-tête TIFF.
3. Dans une nouvelle étape du contrôle, nous vérifions si tous les noms de fichiers sont présents et s'il n'y a pas de doublons. A cette occasion, nous vérifions le nombre de fichiers attendus et comparons si nécessaire la liste des noms de données attendus avec les noms de fichiers effectivement existants²⁴.
4. Dans la prochaine étape, des sommes de contrôle²⁵ sont calculées selon deux procédés différents et consignées dans des listes ASCII. Ces listes ASCII sont tirées sur papier.

²³ Nous utilisons les deux types DLT III/XT et DLT IV.

²⁴ Ces opérations sont basées sur la fonction Unix/Linux `diff` (sous la forme de petits scripts Shell) qui détecte les différences entre les données texte et binaires. Il est à noter que cette même fonction ou des fonctions similaires sont aussi disponibles dans d'autres systèmes d'exploitation. Pour les systèmes d'exploitation Microsoft il existe par exemple Cygnus 32 qui englobe toutes ces fonctions et langages de scripts.

²⁵ Les sommes de contrôle sont calculées avec les fonctions Unix/Linux `cksum` et `md5sum`.

5. Nous copions ensuite les données (avec les sommes de contrôle) sur le/les supports de données avec le "tar". Les supports de données sont ensuite relus et comparés bit par bit avec les données originales²⁶. Dans le même temps, un listing (table des matières) de la bande magnétique est établi. Nous créons en tout trois enregistrements identiques dont deux sont inscrits sur DLT et un sur LTO.
6. Les bandes magnétiques sont ensuite réparties. Un enregistrement est remis au "client", un autre est stocké dans un coffre hors de l'université tandis que le dernier reste à l'université.

5.1.2. Migration

Nous procédons actuellement à la migration des bandes magnétiques DLT vers les bandes LTO. Par rapport au DLT, la technologie LTO offre une plus forte densité de mémoire et est globalement plus moderne. Dans la mesure où la phase de migration actuelle ne porte que sur le changement de support de données mais que le format d'image (TIFF) reste inchangé, l'opération est relativement simple et peut être commandée avec quelques scripts simples²⁷:

1. Une ou plusieurs bandes DLT sont entièrement chargées sur le disque dur.
2. Les sommes de contrôle des fichiers images sont calculées et comparées aux sommes de contrôle antérieures. Si elles coïncident, nous considérons que les fichiers d'images ont été correctement copiés sur le disque dur²⁸.
3. Les données sont ensuite écrites sur LTO et une nouvelle fois relues. Le contenu de plusieurs bandes DLT est écrit sur une bande LTO à cette occasion.

²⁶ Dans le cas du DLT, la bande magnétique est lue avec un autre appareil que celui qui a servi à son écriture. Ce test n'est (malheureusement) pas possible avec les bandes LTO parce que nous ne disposons que d'un seul LTO.

²⁷ Il existe différents langages de scripts sous Linux: Nous utilisons autant les scripts "tsh" que les scripts "Tcl/TK".

²⁸ Pour un niveau de sécurité encore plus élevé, il faudrait charger deux bandes sur le disque dur et une nouvelle fois les comparer bit par bit. C'est à partir de ce moment seulement qu'une sécurité de 100% en matière d'erreurs peut être garantie. La méthode de la somme de contrôle engendre cependant elle aussi une très haute sécurité.

5.2. Schéma type pour le processus de numérisation de matériels photographiques

Le schéma suivant illustre comment un projet de numérisation doit être construit et exécuté. Un catalogue de questions systématique qui fait partie du schéma de processus, doit permettre au lecteur de prendre les bonnes décisions au bon moment. Il importe que le lecteur se rende compte que l'objectif peut être atteint de différentes façons.

- Sélection
 - a. Responsabilités
 - b. Négatif / tirage
 - i. Implications techniques
 - ii. Etat du matériel
 - iii. Considérations supplémentaires
- Copyright
 - a. Responsabilités
 - b. Liste d'adresses pour les sources d'information
- Qui sont les utilisateurs?
- A quelles fins les données numérisées sont-elles utilisées?
- Quelle est la qualité de reproduction requise?
- Sous quelle forme les données doivent-elles être utilisables?
- Aspects conservatoires
 - c. Directives pour la manipulation
 - d. Suggestions pour des modules d'entraînement
- Externaliser la numérisation ou l'effectuer en interne?
 - e. Points à clarifier
- Numérisation
 - f. Paramètres
 - i. Reproduction des tonalités
 - ii. Reproduction des couleurs
 - iii. Définition
 - iv. Bruit
 - v. Choix du matériel
 - vi. Catalogue de questions
 - a. Couleur
 - i. Nuanciers
 - ii. Gestion des couleurs dans une archive numérique
- Métadonnées
 - a. Différentes formes de métadonnées
- Métadonnées techniques
 - b. Que sont les métadonnées techniques?
 - c. Comment les métadonnées techniques sont-elles créées?
- Catalogue de métadonnées NISO
- Métadonnées minimales
 - d. Dublin Core
 - i. Description des éléments minimaux
 - a. Schémas de codage
- XML?

- A quel stade du processus les métadonnées sont-elles insérées?
- Formats de données
 - b. Standards
 - c. Nouveautés
- Produits dérivés
 - d. Utilisation
 - e. Création
- Contrôle qualité (intégré aux étapes nécessaires du schéma de processus)
 - f. Contrôle qualité subjectif
 - g. Contrôle qualité objectif
- Mise en place de la station de contrôle qualité
- Modèles tests avec logiciel d'évaluation
 - h. Quels sont les standards à venir
 - i. Définition
 - ii. Reproduction des tonalités
 - iii. Reproduction des couleurs
 - iv. Bruit